Accounting, Organizations and Society 35 (2010) 565–578

Contents lists available at ScienceDirect

Accounting, Organizations and Society

journal homepage: www.elsevier.com/locate/aos

Financial versus non-financial information: The impact of information organization and presentation in a Balanced Scorecard

Eddy Cardinaels [a], Paula M.G. van Veen-Dirks [b,*]

[a] Department of Accountancy, Tilburg University, The Netherlands
[b] Nijmegen School of Management, Radboud University Nijmegen, P.O. Box 9108, 6500 HK Nijmegen, The Netherlands

A B S T R A C T

This paper investigates how the organization and presentation of performance measures affect how evaluators weight financial and non-financial measures when evaluating performance. We conduct two experiments, in which participants act as senior executives charged with evaluating two business-unit managers. Performance differences between business units are contained in either a financial or one of the three non-financial categories. Specifically, the first experiment studies how organizing measures in a Balanced Scorecard (BSC) format affects performance evaluations. Our results show that when the performance differences are contained in the financial category, evaluators that use a BSC-format place more weight on financial category measures than evaluators using an unformatted scorecard. Conversely, when performance differences are contained in the non-financial categories, whether measures are organized into a BSC-format or into an unformatted scorecard has no impact on the evaluation. The second experiment shows that when performance markers are added to the scorecards (i.e., +, −, and = signs for above-target, below-target, and on-target performance), evaluators that use a BSC-format weight measures in any category containing a performance difference more heavily than evaluators using an unformatted scorecard. Our findings suggest that firms should carefully consider how to present and organize measures to get the intended effect on performance evaluations.

Introduction

Kaplan and Norton (1992) originally introduced the Balanced Scorecard (BSC) to overcome problems that result from a sole focus on financial measures. A BSC enables financial performance measures (grouped into a single financial category) and non-financial performance measures (grouped into non-financial categories including customer, internal business process, and learning and growth) to be displayed in combination. In practice, the format of performance scorecards varies significantly across firms (Lohman, Fortuin, & Wouters, 2004). Some firms organize their measures into BSC performance categories while others simply provide a general list of measures. How results are presented in a scorecard also varies. Many firms show only target levels and actual results, while other firms supplement this information with performance markers (i.e., +, −, =) or qualitative signs (e.g., red, yellow, and green indicators) to more explicitly indicate the status of the actual results in relation to the target levels (e.g., Malina, Norreklit, & Selto, 2007; Malina & Selto, 2001; Merchant & Van der Stede, 2007). Despite the prevalence of these different formats in practice, little work has been done on how variations in scorecard formats affect performance evaluations.

In this study, we examine how variations in, first, the organization (i.e., BSC versus unformatted scorecard) and, second, the presentation of measures (i.e., the use of

* Corresponding author.
E-mail addresses: e.cardinaels@uvt.nl (E. Cardinaels), p.vanveen@fm.ru.nl (P.M.G. van Veen-Dirks).

0361-3682/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aos.2010.05.003

markers) affect how evaluators weight financial and non-financial measures in performance evaluations. Prior studies have primarily focused on the finding that, when firms use both common measures (i.e., measures common across multiple units) and unique measures (i.e., measures unique to particular units) for their business units, evaluators ignore the unique measures (Lipe & Salterio, 2000). Solutions to this problem have also been explored (Libby, Salterio, & Webb, 2004; Roberts, Albright, & Hibbets, 2004). Many firms, however, use similar scorecards that contain only measures common to all business units (e.g., Malina & Selto, 2001). In such cases, presentation formats and features may well affect how evaluators weight financial and non-financial information in performance evaluations. To investigate these issues, we present two experiments that extend the basic setup of Lipe and Salterio (2002).

Lipe and Salterio (2002) study how information organization (i.e., how organizing measures into a BSC as opposed to an unformatted list) affects the performance evaluation of two business-unit managers. They consider, however, only the case wherein performance differences between the two business units (i.e., consistent above-target performance for one business unit and consistent below-target performance for the other) are located on the non-financial category of customer measures. They show that evaluators using a BSC weigh these measures less heavily than evaluators viewing the same measures in an unformatted scorecard.

Our first experiment extends Lipe and Salterio's work by examining whether the effect of how the measures are organized depends on which type of category—that is, financial or non-financial—contains the performance differences between business units. We predict that information organization will have a greater effect on evaluations when performance differences appear in the financial category. We base this prediction on performance-measurement as well as psychology literature, which suggest both that people are heavily led by financial outcomes and that how people use a BSC to process information may lead these users to place more weight on financial performance measures than users of an unformatted scorecard. We use a 2×4 design, manipulating how information is organized (i.e., in a BSC or an unformatted scorecard) when performance differences between two business units are located in either the financial category or one of three non-financial categories. We qualify the results of Lipe and Salterio (2002) by showing that a BSC only "increases" the weight evaluators attach to performance differences when these differences are located in the financial category. We find that when performance differences are located in one of the three non-financial categories, information organization has no effect. We thus also observe no decrease in how measures are weighted for the customer category, which is the only case comparable to that of Lipe and Salterio (2002). We attribute this latter finding to some differences in design choices, which we will explain in 'Methods and results'.

Increasing the weight evaluators place on financials may not always be the effect firms hope to achieve by using a BSC instead of an unformatted list of measures.[1] Therefore, our second experiment examines whether the use of markers (i.e., +, −, and = signs for above-target, below-target, or on-target performance) offers a counterbalancing effect. The design of Experiment 2 is similar to that of Experiment 1 except that we add performance markers to the scorecards' results. We hypothesize, and find, that, when supplemented with markers, performance differences on measures of any category, be it financial or non-financial, are always weighted more heavily in a BSC than in an unformatted scorecard.

Our research contributes to the literature in several ways. First, prior results on the use of financial and non-financial measures are still inconclusive (Luft & Shields, 2001; Schiff & Hoffman, 1996). Although the BSC has gained prominence in accounting research as a way of integrating financial and non-financial performance measures (Hoque & James, 2000), we show a consequence of organizing the measures into the BSC categories that may well be uncalled-for if firms adopt a BSC to stimulate the use of non-financials. Our finding in Experiment 1 that a BSC only increases the weight evaluators assign to the financial category, leaving non-financial categories unaffected, adds a new issue to the BSC literature, which to date has focused on the problem of common versus unique measures.

Second, we show how different presentation formats can produce different processing strategies (Payne, 1982; Schkade & Kleinmuntz, 1994). In Experiment 1, we show that grouping and labeling measures (i.e., in a BSC), as opposed to leaving measures unlabeled and in no particular order (i.e., in an unformatted scorecard), helps evaluators identify financials more easily and may activate their beliefs in the relative importance of financials. As a result, a BSC-format increases an evaluator's basic tendency to weight financial measures more heavily than non-financial measures. Experiment 2 shows that performance markers in a BSC can also direct an evaluator's attention to other non-financial categories that contain important performance differences. In this case, BSC users compared with users of an unformatted scorecard, give more weight to any category (financial and non-financial alike) that shows consistently good or bad performance.

These findings have important practical implications for the many firms that use the BSC as a tool to evaluate and reward managers (Kaplan & Norton, 1996; Liedka, Church, & Ray, 2008, p. 74). If evaluators assimilated all measures without bias, then the format of a scorecard would not matter. However, because format, in fact, appears to have a strong impact on how evaluators assimilate measures, firms should carefully consider how they display these

[1] We investigate how organization and presentation of measures affect how evaluators subjectively weigh performance differences on either financial or non-financial measures (when a firm uses a common set of measures). Exact weights often cannot be prescribed because they depend on many factors, including the sensitivity, precision, congruency, and quality of the measures (Libby et al., 2004). We therefore avoid the term *bias*. However, if only the financial measures impact performance evaluations, and non-financials have no impact, we question whether this is always in line with the idea of using a BSC (as opposed to an unformatted scorecard) to direct an evaluator's attention toward the firm's non-financials.

measures. Given that managers' behavior is driven by weights placed on the performance measures (e.g., Ittner, Larcker, & Meyer, 2003; Smith, 2002), formatting can thus have far-reaching consequences for the firm.

Hypothesis development

Assessing and combining the scores of various performance measures into an overall evaluation is a complex task (Lipe & Salterio, 2000, 2002). Due to information processing limitations (Baddeley, 1994; Hastie, 1991; Shanteau, 1988), evaluators often have cognitive difficulties making evaluation judgments. While Kaplan and Norton (1996) have proposed the BSC as a tool that enables managers to utilize important non-financial as well as financial measures, prior work has not yet clearly shown how a BSC, as an information-organization device, affects evaluators' cognitive processes and their use of financial and non-financial performance information in evaluations.

Many studies on the BSC have experimentally examined the differences in weighting between common and unique measures (Banker, Chang, & Pizzini, 2004; Libby et al., 2004; Lipe & Salterio, 2000). When a BSC uses both common measures (i.e., measures common across business units) and unique measures (i.e., measures specific to one business unit), evaluators place more weight on common measures, while ignoring unique measures that may also be informative (Lipe & Salterio, 2000). Only a few authors (e.g., Lipe & Salterio, 2002) have studied subtle presentation effects, for example, how the categories used in a BSC impact an evaluator's judgment. Studying presentation effects, however, is important because information organization and presentation can affect an evaluator's processing strategies (Kleinmuntz & Schkade, 1993; Payne, 1982) and, in turn, his or her use of financial or non-financial information. Moreover, many firms use only a common set of financial and non-financial measures to evaluate their business-unit managers. Guided by the producers of performance-measurement software, who offer packages for monitoring business performance, firms organize and present performance information in various ways. Status indicators, for both variances between target and actual performance and trends in performance, often accompany performance results (e.g., www.sap.org; www.ibm.org). Some firms also design their own scorecards. For example, General Electric's "digital cockpit" provides a tabular summary of performance, complemented by color-coded indicators for each measure's status (Few, 2006).

We explicitly study how the organization and presentation of measures impact how evaluators weight financial and non-financial information in performance evaluations. Hypothesis 1—which we test in Experiment 1—predicts how organizing information into a BSC versus an unformatted scorecard affects the weights evaluators attach to financial and non-financial measures. Lipe and Salterio (2002) examined the effect of information organization when performance differences were located in the customer perspective. They argued that when measures are grouped together and show consistently good or bad performance, BSC users perceive them as being more related, and, in turn, give these seemingly related items less weight than users of a format in which the same items are placed in no particular order. Lipe and Salterio (2002) assumed that their prediction would hold for all BSC categories, both financial and non-financial. We, however, predict the opposite effect when performance differences are located on financial measures. Our theory depends on evaluators' basic tendency to rely on financial measures—a tendency that, we predict, will manifest more strongly with a BSC than with an unformatted scorecard.

Hypothesis 2—which we test in Experiment 2—addresses how measures are presented. We predict the effect of information organization (BSC versus an unformatted scorecard) when both scorecards contain performance markers. We argue that, compared with the use of markers in an unformatted scorecard, the use of markers in a BSC helps evaluators to pursue the strategy to rely heavily on the perspectives with consistent performance differences across business units. Hence, we predict that all categories, non-financial and financial alike, that demonstrate consistent performance differences across business units will be given more weight when presented in a BSC than in an unformatted scorecard. As such, markers in a BSC can be a useful tool for directing attention toward non-financials. The following text develops these hypotheses.

Organization of measures and the weighting of financial and non-financial information

In this section, we argue that how information is organized can reinforce an evaluator's tendency to rely on financial measures such that, users of a BSC compared with users of an unformatted scorecard, will weight consistent performance differences on financial measures more heavily than consistent performance differences on non-financial measures.

Evidence suggests that managers tend to weight financial measures more heavily than non-financial measures for reasons including outcome effects, outside pressure, and familiarity. The psychology literature argues that evaluators are susceptible to the "outcome effect" (Mitchell & Kalb, 1981), which states that, when assessing a manager's performance, evaluators give outcomes more weight in their evaluations (Ghosh & Lusch, 2000; Hawkins & Hastie, 1990), regardless of whether the actions to achieve the results were appropriate (Ittner et al., 2003). Typically, financial measures (e.g., sales growth, sales margins) share a common orientation toward financial outcomes whereas non-financial measures contain a mixture of outcome-oriented measures and measures seen as drivers of such outcomes (e.g., returns to suppliers, retail experience of employees). Empirical work also suggests that people are familiar with companies' financial pressures because shareholders are vocal and boards frequently apply pressure on behalf of shareholders (Anthony & Govindarajan, 2001). DeBusk, Brown, and Killough (2003) believe that managers rely on those measures with which they are most familiar, that is, financial measures. This may, in turn, reinforce the tendency to rely on financial measures. Indeed, this strong reliance on financial measures seems to

occur in practice (DeBusk et al., 2003; Ittner & Larcker, 1998; Ittner et al., 2003).[2]

Multiple reasons exist for why grouping financial measures together and labeling them financial, as in a BSC, rather than mixing them with non-financial measures and leaving them unlabeled, as in an unformatted scorecard, increases evaluators' tendency to weight financial measures more heavily than non-financial ones. First, it is cognitively difficult for evaluators to assess differences between actual and target results for large sets of measures (Payne, 1982). Lipe and Salterio (2002) have argued that a BSC can help evaluators to mentally organize a large number of performance measures. The BSC divides measures into smaller groups of performance categories, which allows subjects to mentally invoke a "divide and conquer" or group-based processing strategy (Lipe & Salterio, 2002; Shanteau, 1988). Rather than processing all measures simultaneously, evaluators assess the measures by group before combining them into an overall judgment. Assessing measures by group, as is possible with a BSC, is likely to be less cognitively demanding (Kaplan & Wisner, 2009; Lipe & Salterio, 2002). As such, grouping and labeling measures ensures that evaluators using a BSC will, in fact, identify financial measures as such and separate them from non-financial measures (Koonce et al., 2005; Maines & McDaniel, 2000). In contrast, users of the unformatted scorecard, still have to select among a large unordered set of measures (Payne, 1982), whereby it remains cognitively difficult to assess all the relevant financial measures. Given that evaluators have a tendency to rely on financial outcomes and that financials are easier to identify in a BSC than in an unformatted scorecard, BSC users are more likely to thoroughly assess the financial measures as a group and, consequently, to give financials more weight when information of different categories is being combined.

Second, grouping and labeling also suggest that the distinction between financial and non-financial matters. It is often argued that presentation formats and labels can frame the decision into a certain context and influence decision makers to make different judgments (e.g., Maines & McDaniel, 2000; Vera-Muñoz, Kinney, & Bonner, 2001). The labels provided in a BSC may cue evaluators to activate their beliefs about the relative importance of financial measures—that outcomes matter, that outside stakeholders care about financial performance, etc.—while these beliefs are less likely to be activated by the unformatted scorecard.

Based on the above, we expect that the organization of information (i.e., in a BSC versus an unformatted scorecard) will produce strong differences in how measures are weighted when performance differences between business units are located in financial, rather than non-financial, categories. Due to the group-based processing strategy (Shanteau, 1988), evaluators using a BSC are more likely to identify financial performance as consistently superior for one business when they assess the financials as a group. As such, when information from different categories is combined, evaluators using a BSC may give a performance difference on financials more weight in the overall evaluation. Second, due to labeling, cues about the importance of financials are also likely to be activated when financials strongly point in the same direction. Because of these two effects, the judgment of financial performance differences becomes relatively extreme with a BSC. This superior performance on financials is less apparent in an unformatted scorecard, which does not allow processing at the group level. Instead, evaluators must deal with a large set of unordered measures, which is cognitively more difficult (Payne, 1982). They may assess only a limited set of measures (Payne, Bettman, & Luce, 1998), which may not include all financial measures. Moreover, because the label financial is absent, identifying the financial measures as such and separating them from the non-financials is not as easy. Hence, users of an unformatted scorecard are likely to make less extreme evaluations than users of a BSC.[3]

If performance differences are located on the non-financials, people may still focus heavily on financial outcomes. Because of this focus on financials and the fact that group-based processing in a BSC makes it easier to identify the financial measures, BSC users are likely to make a thorough assessment of the financial measures. As such the absence of differences on the 'financials' might still heavily influence the overall judgment of a BSC user, even though one of the non-financial categories contains the performance differences. The lack of specific differences on the financials thus makes the evaluation less extreme. Again, assessing performance measures at the group level is less obvious in an unformatted scorecard. Instead these evalu-

[2] A 1996 Towers Perrin survey that found BSC adopters were willing to place, on average, 56% of the relative weight on financial measures provides such evidence (Ittner and Larcker, 1998). DeBusk, Brown, and Killough (2003) also found that users of performance measurement systems view bottom-line financial measures as more important than non-financial measures. In their case study, Ittner et al. (2003) further found that, when determining employee bonuses, evaluators place the most weight on quantitative, outcome-oriented financial measures (p. 754). Ittner et al. (2003) further note that, in bonus plans, evaluators ignored many leading (non-financial) indicators for firm performance.

[3] We assume that users of an unformatted scorecard compared to users of a BSC, may not select all financial measures that contain a significant performance difference, because of the cognitive difficulties of having to select among a large set of measures. Note that even if we assume that people with an unformatted scorecard would select the same set of measures as people with a BSC, the classification that participants have to make in an unformatted scorecard, relative to the classification that is given to BSC-users, may still explain the differences we obtain. Assume that performance differences are located on financials. Purely because of the labeling of measures as financials (Koonce, Lipe, & McAnnally, 2005), participants with a BSC may accept all the measures in this category as financials and give it more weight. In an unformatted scorecard, participants using the same measures may still label some of these measures as non-financial. Inventory turnover, for example, is a financial measure in our BSC, while users in an unformatted scorecard might perceive it as a non-financial measure. If we assume that users in general give less weight to non-financials (because they focus heavily on financial outcomes), users of an unformatted scorecard might give the same set of measures less weight because they consider some of these measures as non-financials. Yet, our subsequent tests of the measures that participants have used in their judgment, suggest that users of an unformatted scorecard compared to users of an unformatted scorecard, consider different measures to be more important, suggesting that they indeed make a different selection of measures (as we have argued).

ators have to select among a large unordered set of measures, which might lead them to select only a few financial measures—which show no specific differences—and only a few non-financial measures—which show important differences. As argued before, this can also make the judgment less extreme. Because judgments are less extreme under both types of scorecards we expect that information organization has less effect in case of a performance difference on a non-financial category. In sum, for Experiment 1, we predict an interaction effect suggesting that the weighting of financial measures compared to non-financial measures will depend on information organization:

H1. The use of a BSC, compared with the use of an unformatted scorecard, increases an evaluator's basic tendency to weight financials more heavily than non-financials.

The above also provides an alternative explanation for the results of Lipe and Salterio (2002). Their finding that, compared with users of an unformatted scorecard, BSC users assigned less weight to customer-related measures may simply have been because these BSC users gave more weight to financial outcomes in their overall evaluations. Unlike in our study, the financial category in Lipe and Salterio (2002) showed a slightly positive performance for both business units, when performance differences are contained in the customer perspective. As such, the reduced weighting of customer measures might not be caused by perceived correlations of these measures resulting from their being grouped together in a BSC, as Lipe and Salterio (2002) argue. Rather, when BSC users engage in the process of combining information of different categories into an overall evaluation, the fact that both units score equally well on financial outcomes can make the difference in judgment less extreme.

Presentation of measures and the weighting of financial and non-financial information

Performance information can be visually represented using attributes such as location, color, length, and size, highlighting patterns and trends that might otherwise not be visible (Card, Mackinlay, & Shneiderman, 1999). In performance scorecards, status indicators are often used. Azofra, Prietro, and Santidrian (2003), for instance, report in a case study that the control instrument uses traffic-light colors to highlight the status of the indicators. In their study of best practice in performance management, Bauer, Tanner, and Neely (2004) find that using such traffic-light reporting was common practice. This section explores the differences between a BSC and an unformatted scorecard when scorecards contain performance markers (i.e., explicit +, −, and = signs for above-target, below-target, and on-target performance). We predict that, when scorecards contain markers, BSC users, compared with users of unformatted scorecards, will give more weight to any type of category, be it financial or non-financial, containing a consistent performance difference.

If evaluators assessed all performance cues without bias, they would compare the actual results with the target results for all measures, and adding extra presentation features would not make a difference (Haynes & Kachelmeier, 1998). Yet, as we have argued, comparing actual and target results of a large set of measures is cognitively difficult. Supplementing the information with performance markers may still facilitate this information-processing task (Kleinmuntz & Schkade, 1993; Libby, 1981; Schkade & Kleinmuntz, 1994; Silver, 1991). Performance markers enable evaluators to view the differences between actual and target results for all measures on a scorecard at a glance. Not having to compare actual and target results for each individual measure can save them considerable cognitive effort (Ganzach, 1994). Moreover, performance markers can make informational items stand out relative to other stimuli in the environment and can thereby redirect evaluators' attention to such items (Almer, Hopper, & Kaplan, 2003; Fiske & Taylor, 2008; Haynes & Kachelmeier, 1998).

This is particularly true when performance markers are used in a BSC, wherein any category containing consistent performance differences (i.e., consistently above or below-target performance) will stand out relative to other categories. Performance markers in a BSC indicate that all the measures in a category containing a consistent performance difference have similar values (i.e., all + or all −). Without performance markers, BSC users would focus heavily on a thorough assessment of financial measures and, as such, effects on non-financials may not always be fully accounted for. Because of the saliency effect of performance markers (Almer et al., 2003), BSC users can, at a glance, fully assess the performance differences in each of the four categories, and thereby give more weight to information in the category containing a consistent performance difference relative to others that show no specific difference. Indeed, focusing on the category containing a consistent performance difference can be an important strategy for processing performance information, and presentation formats, like performance markers, can make this strategy more accessible to evaluators (Ganzach, 1994; Sundstrom, 1987). Especially, when performance markers show systematic performance differences between business units (Kulik, 1989), evaluators can become more extreme in their judgments (Ganzach, 1994). As a result BSC users tend to weight the category containing a consistent performance difference more heavily in their overall evaluation.

Adding performance markers to an unformatted scorecard also saves time in that performance differences do not have to be assessed on each measure. Evaluators, however, would still have to combine the scores for all available measures into an overall evaluation, which is a cognitively challenging task (Payne, 1982). As argued, processing performance information at the group level is much more difficult for users of an unformatted scorecard because an unformatted list of measures does not group items into labeled categories. It is therefore difficult for evaluators to establish that one business unit has indeed consistently outperformed the other on a specific dimension, and their evaluations may be less extreme than those of BSC users. As a result, users of an unformatted scorecard compared to BSC users attach less weight to the measures from the category showing consistent performance differences. Hence, for Experiment 2, we predict a main effect of information organization:

H2. The use of a BSC with markers, compared with the use of an unformatted scorecard with markers, increases the weights evaluators place on both financial and non-financial measures.

Methods and results

Selection of the performance measures for both experiments

For both experiments, we use case materials adapted from prior studies on the BSC (e.g., Banker et al., 2004; Lipe & Salterio, 2000, 2002). Participants assume the role of senior executive of the retail firm, "VQS Inc.," which specializes in clothing. Participants review the performance of two VQS business units, "Streetware" and "Family Fashion." Streetware specializes in youth fashion, and Family Fashion in clothing for young families. Managers and strategies for these two business units are described in detail. As in Lipe and Salterio (2002), participants in both experiments are explicitly told that the performance metrics are appropriate for retailers and capture the various aspects of each business unit's strategy. For each business unit, we used a set of 16 common measures, with four per category. Given our interest in how participants assess similar performance differences based on which BSC category contains those differences, it is important that we (1) select measures perceived as being typical for the BSC category in question and (2) that how typical these measures are does not significantly vary across categories.

To be sure we satisfied these two requirements, we first performed a pilot test. We drew our set of measures from Lipe and Salterio (2000) and Banker et al. (2004), both of which used a range of 24 measures. In the pilot test, 54 students reviewed the measures of Lipe and Salterio (2000) and an additional 46 students reviewed those of Banker et al. (2004). The students assessed, on a 10-point scale, how "typical" each measure was for its BSC category (with 1 indicating "not typical at all" and 10 "very typical"). The mean rating of all measures was 6.8. The measures we retained—four per category—had mean ratings of 7.15, 7.23, 7.10, and 7.13 for the financial, customer, internal business process, and learning and growth categories, respectively. These means were not significantly different from each other (for all comparisons, $p > 0.22$, two-tailed) and all fell slightly above the overall mean of 6.8 (all p's < 0.05, two-tailed). Table 1 presents the 16 measures we retained. Given that our measures scored above average (in terms of their typicality) and that this score did not vary significantly across categories,[4] we can assume that our results are not driven by one category's measures seeming less typical than those of another category.

Because of this pilot test our study uses a different set of performance measures than Lipe and Salterio (2002). Another important difference between our study and that of Lipe and Salterio (2002) is that, in our study, any category containing no performance differences between the two business units exhibits no specific trend (i.e., categories always contained one above-target measure and one below-target measure). While this is mostly the case in Lipe and Salterio (2002), their financial category does show a positive performance for both business units (i.e., there are two above-target measures and only one below-target measure). As argued, evaluators using a BSC in Lipe and Salterio (2002) may have given less weight to the customer measures simply because both business units performed well on the financial measures.

Experiment 1

Experiment 1 extends the work of Lipe and Salterio (2002) by studying the effect of information organization when performance differences between the two business units are shifted across the four types of measures of a BSC. It tests H1, which states that organizing measures in a BSC, as opposed to in an unformatted scorecard, increases an evaluator's tendency to weight financial measures more heavily than non-financial measures.

Experimental manipulations, participants, and procedures

In Experiment 1, we use a 2×4 between-subjects design. We manipulate the type of scorecard (i.e., a BSC versus an unformatted scorecard) as well as the category of measures (i.e., financial, customer, internal business, or learning and growth) containing the performance differences between the two business units. We further counterbalance the order in which participants evaluate Streetware and Family Fashion.

Consistent with Lipe and Salterio (2002), the factor 'organization' has two levels. As shown in Table 1, a BSC organizes the 16 measures into the four perspectives. Participants in the unformatted scorecard condition receive the same 16 measures in no particular order in an unlabeled list. The order of the measures in this list was randomly fixed: of the 16 possible positions, the financial measures were on positions 3, 5, 10, and 16; the customer measures on positions 2, 8, 9, and 15; the internal-process measures on positions 4, 6, 11, and 13; and the learning and growth measures on positions 1, 7, 12, and 14.[5]

Our 'type of measure' manipulation has four levels. The same degree of good (or poor) performance by a business unit is situated in the financial, customer, internal business, or learning and growth category. By shifting the same excellent performance across the four types of BSC categories, we extend the work of Lipe and Salterio (2002), who studied the effect of information organization only for the case in which one business unit outperformed the other on customer measures. Table 1 provides more detail on this manipulation. The first column of actual measures

[4] An item in the post questionnaire revealed that participants in both experiments perceived the selected measures as relevant for the two business units (on a 7-point likert scale with 1 equal to disagree and 7 to agree). The mean of 5.13 suggests that participants perceived the measures as relevant (t-test different from 4, $t = 17.10$, $p < 0.001$). Importantly, the scores did not significantly differ for the between-subject factors "type of measure" ($F = 1.38$, $p = 0.25$), "organization" ($F = 0.61$, $p = 0.44$) and "order" ($F = 0.17$; $p = 0.68$). There were also no differences in the perceived relevance of the measures between the two experimental groups that received the scorecards with and without performance markers ($F = 0.41$; $p = 0.52$).

[5] In line with Lipe and Salterio (2002), we use a blank line after every four measures so that eye fatigue and readability did not vary between the two formats.

Table 1
Type of measure manipulation (between-subjects factor).

Measures and targets for Streetware (Family Fashion in brackets)	Streetware excels (Family Fashion performs poorly) on all four measures			
	Target	Financial measures Actual	Customer measures Actual	Internal measures Actual	L&G measures Actual
Financial					
Sales margins (%)	60.0 (62.0)	66.0 (55.8)	66.0 (68.2)	66.0 (68.2)	66.0 (68.2)
Sales growth per store (%)	15.0 (18.0)	15.8 (17.1)	15.0 (18.0)	15.0 (18.0)	15.0 (18.0)
Inventory turnover	6.0 (5.0)	6.6 (4.5)	5.4 (4.5)	5.4 (4.5)	5.4 (4.5)
Percentage of sales from new stores (%)	30.0 (25.0)	31.5 (23.8)	30.0 (25.0)	30.0 (25.0)	30.0 (25.0)
Customer					
Customer satisfaction rating (%)	85.0 (90.0)	93.5 (99.0)	93.5 (81.0)	93.5 (99.0)	93.5 (99.0)
Sales per square foot of retail space	30,000 (25,000)	30,000 (25,000)	31,500 (23,750)	30,000 (25,000)	30,000 (25,000)
Repeat sales (%)	30.0 (40.0)	27.0 (36.0)	33.0 (36.0)	27 (36.0)	27.0 (36.0)
# of new items in which first to market	70.0 (60.0)	70.0 (60.0)	73.5 (57.0)	70.0 (60.0)	70.0 (60.0)
Internal business processes					
Returns to suppliers (%)	6.0 (4.0)	5.4 (3.6)	5.4 (3.6)	5.4 (4.4)	5.4 (3.6)
Average markdowns (%)	15.0 (12.0)	15.0 (12.0)	15.0 (12.0)	14.3 (12.6)	15.0 (12.0)
Orders filled within one week	3000 (2500)	2700 (2250)	2700 (2250)	3300 (2250)	3300 (2250)
# of stock-outs	2.0 (3.0)	2.0 (3.0)	2.0 (3.0)	1.9 (3.2)	2.0 (3.0)
Learning and growth					
Hours of sales training per employee	15.0 (13.0)	16.5 (14.3)	16.5 (14.3)	16.5 (14.3)	16.5 (11.7)
Suggestions per employee	1.0 (2.0)	0.9 (1.8)	0.9 (1.8)	0.9 (1.8)	1.1 (1.9)
Retail experience of sales managers	3.0 (4.0)	3.0 (4.0)	3.0 (4.0)	3.0 (4.0)	3.3 (3.6)
Employee satisfaction (%)	80.0 (82.0)	80.0 (82.0)	80.0 (82.0)	80.0 (82.0)	84.0 (77.9)

The table shows the type of measure manipulation. The measures showing excellent performance for Streetware (poor performance for Family Fashion) had two +5% and two +10% above-target measures (two −5% and two −10% below-target measures). Of the three categories that showed no performance differences between Streetware and Family Fashion, two measures were on target, one measure was +10% above target and one measure −10% below target. Organization is our second between subject manipulation, that is, whether the measures were presented in a BSC (see Table 1) or in an unformatted scorecard containing the same measures in no particular order. We also counterbalanced the order in which participants evaluate Streetware and Family Fashion. In Experiment 1, we did not add any presentation effects. In Experiment 2 we add performance markers to the scorecards (+, −, and = signs for above-, below- or on-target performance).

presents the condition in which performance differences are located on the financial measures: Streetware performs above-target on all four financial measures (twice 5% and twice 10% above-target), whereas Family Fashion performs below-target on all four financial measures (twice 5% and twice 10% below-target). The business units show no specific difference in performance on the remaining categories of measures (both units have a 10% above-target, a 10% below-target, and two on-target realizations). The second, third, and fourth columns present the remaining three conditions in which performance differences are located on each of the three non-financial categories, that is, either on the customer, internal business, or learning and growth category.[6]

We recruited 144 students from a 4-year business program (comparable to study at the master level) at a large West European university via accounting courses scheduled in the final 2 years of their curriculum. Through such core managerial accounting courses, these students were familiar with the concept of a BSC and its use as a tool for measuring the performance of business units. They had an average of 3.84 years work experience acquired via part-time jobs and internships in retailing (58%), other industries (69%), accounting or auditing (25%), and marketing (15%). Sixty-eight percent of our sample were male and most participants (97.2%) indicated that they had visited a retail clothing store in the past 12 months.[7] We administered the experiment by computer, and participants were randomly assigned to one of the experimental treatments. After reading the case descriptions, participants were asked to evaluate the performance of each of the two business-unit managers, on a scale from 0 to 100, using seven descriptive labels, as was the case in Lipe and Salterio (2000, 2002).[8] We also asked additional questions on these evaluations (e.g., what type of measures the participant had used). Each evaluation was performed with the scorecard of the respective

[6] When administrating the cases, we randomized which two measures were 5% above (5% below) and which two were 10% above (10% below) target for the "type of measures" containing the good (bad) performance. For measures of other BSC categories, we again used randomization to set one measure 10% above target, one measure 10% below target, and the remaining two measures on target.

[7] Our participants are reasonably comparable to the MBA students that were used in Lipe and Salterio (2002). All participants have covered the basics of a BSC, as one would in a core managerial accounting MBA course. Also, like participants in Lipe and Salterio (2002), who had only 4 years of work experience (i.e., the equivalent of entry-level managers), our participants acquired 3.84 years of part-time work experience (i.e., again, comparable to that of entry-level management) through internships and part-time jobs.

[8] The labels used were excellent: far beyond expectations, manager excels; very good: considerably above expectations; good: somewhat above expectations, average: meets expectations; poor: somewhat below expectations, needs some improvement; very poor: considerably below expectations, needs considerable improvement; and reassign: sufficient improvement unlikely.

business unit displayed on the computer screen. We ended Experiment 1 with a questionnaire containing items on task understanding, realism, and motivation, followed by some manipulation checks. Each session lasted about an hour, and participants received course credit for their participation.

Results

We focus on the difference in evaluation scores (Banker et al., 2004; Lipe & Salterio, 2002) between Streetware and Family Fashion to asses how much weight evaluators give to performance differences located on financials and non-financials and how organization (BSC compared to an unformatted scorecard) affects this weighting (H1).[9] Table 2 shows the mean differences in evaluations for the experimental cells in Experiment 1. Because our type of measure manipulation has four levels, we use contrast analyses to analyze H1 (e.g., Buckless & Ravenscroft, 1990). Given our prediction in H1, we always contrast the cells with performance differences located in the financial measures against those cells with performance differences located in the non-financial measures.[10]

The means in Panel A of Table 2 show that the BSC compared to the unformatted scorecard has a greater effect on evaluations when performance differences are located on financial measures. Conversely, when one business unit outperforms the other on non-financial measures, we find no significant difference in weighting between the two formats. As shown in Table 2, Panel B, the interaction of financial versus non-financial information by organization is significant. Consistent with H1, how financial versus non-financial information is weighted indeed depends on how that information is organized. To be conservative, we report two-tailed statistics in our tables. Nevertheless, given that H1 is directional, we could argue that the effect size of this interaction (i.e., +10.167, $t = 1.72$, $p = 0.0434$, one-tailed) is also significant at the 5% level.

Consistent with H1, the findings in Panel C further confirm that the BSC-format (as opposed to an unformatted scorecard) increases an evaluator's tendency to weight financial measures more heavily than non-financial measures. The effect of information organization on financials is equal to +10.33 ($p = 0.045$). Conversely, the BSC-format has no effect on how non-financials are weighted (the effect of information organization for non-financials = +0.17, $p = 0.955$). Panel C further shows that when measures are organized into a BSC, financial measures are weighted more heavily than non-financial measures (+13.02, $p = 0.002$). This effect is not significant in an unformatted scorecard (+2.85, $p = 0.496$).[11] In sum, our results imply that grouping and labeling multiple measures under the four BSC perspectives—as opposed to arranging the measures randomly—does not help those firms that desire to stimulate evaluators' use of non-financial measures.[12]

When studying the effect of information organization for each category of non-financial measures (untabulated results), we find that performance differences on customer, internal business, and learning and growth measures are weighted no differently with a BSC than with an unformatted scorecard (i.e., the effect of organization on (1) customer measures: +0.50, $p = 0.923$; (2) internal business −0.33, $p = 0.949$; and (3) learning and growth measures: +0.33, $p = 0.949$). Table 2, Panel A supports these results. Comparing our results in the customer category with those of Lipe and Salterio (2002), who studied performance differences in this category only, suggests an important contradiction. While Lipe and Salterio (2002) found that evaluators gave less weight to customer measures under the BSC than under the unformatted scorecard, we observe no difference in the weighting of these measures as a result of information organization. This contradiction can be explained in several ways. As mentioned in 'Methods and results', we used a different set of measures than those used in Lipe and Salterio (2002). Second, in Lipe and Salterio (2002), when the customer category contained the performance differences, the financial category also showed a slightly positive trend for both business units. Consistent with H1, the fact that both units scored well on financials may have made the evaluation judgments of participants using a BSC-format less extreme and, as such, they may have reduced the weights they assigned to the customer category.

Supplementary analyses

After our participants had completed their evaluations, we asked them to list, for each business unit and in decreasing order of importance, the first five measures

[9] We also analyzed whether the BSC-format versus the unformatted scorecard had an effect on the variability of performance. Yet, std. dev. did not differ significantly between the two groups (p-value of the Levene's test = 0.23). Furthermore, std. dev. of cell means on financials (13.18 versus 17.57, p-value of the Levene's test = 0.23) and std. dev. of cell means on non-financials (16.24 versus 17.13, p-value Levene's test = 0.99) did not differ significantly between BSC-format and the unformatted scorecard.

[10] As discussed further, we also report individual tests on how organization affects the weighting on each type of the non-financial measure manipulation (i.e. performance differences either located on customer, internal business or learning and growth measures). In particular, results for the experimental cells in which performance differences are located on the customer perspective can directly be compared against those in Lipe and Salterio (2002), who merely focused on performance differences in this specific category.

[11] We analyzed score differences in greater detail by examining the individual business-unit level. We do not observe any particular differences for Streetware. The results are fully driven by Family Fashion (which shows poor performance). In particular, when performance differences are located in the financial category, BSC users evaluate Family Fashion significantly lower than do users of the unformatted scorecard (54.11 versus 45.00, $p = 0.04$). When performance differences are located in one of the non-financial categories, no significant differences exist between how BSC users and users of an unformatted scorecard evaluate Family Fashion (54.67 versus 55.43, $p = 0.77$). In particular, poor "financial" performance is weighted more heavily in a BSC than in an unformatted scorecard.

[12] The stronger weighting of financials is not caused by the fact that a BSC lists the financial measures first. We ran an additional test with 14 students, in which the learning and growth category of the BSC was the first category listed. This category contained the performance differences between the two business units. As we observe in Table 2, evaluators still ignore these performance differences; the evaluation difference between Streetware and Family Fashion was only 6.42. This was not significantly different from the mean of 4.78 in Table 2 ($p = 0.75$) when performance differences were located in the L&G category (but with the financial category listed first). Therefore, putting a category with performance differences on top of a BSC does not increase its weighting (i.e., information that is ignored remains ignored, even when listed first in a BSC).

Table 2
Results of Experiment 1 (test of H1).

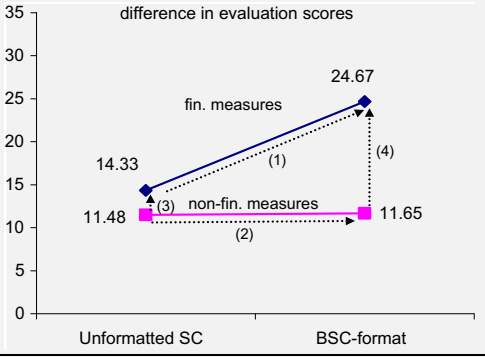
Panel A: Summary statistics per experimental cell[a]

Scorecard organization	Type of measure				
	Financial measures	Customer	Internal	L&G	Non-financial measures
Unformatted Scorecard					
Eval. Streetware	68.44	70.44	66.50	61.50	66.15
Eval. Family Fashion	54.11	49.11	57.83	57.06	54.67
Difference in eval.	14.33	21.33	8.67	4.44	11.48
	[13.18]	[13.90]	[12.07]	[17.89]	[16.24]
	(n = 18)	(n = 18)	(n = 18)	(n = 18)	(n = 54)
BSC-format					
Eval. Streetware	69.67	72.06	66.39	62.78	67.08
Eval. Family Fashion	45.00	50.22	58.06	58.00	55.43
Difference in eval.	24.67	21.83	8.33	4.78	11.65
	[17.57]	[16.01]	[18.78]	[11.60]	[17.13]
	(n = 18)	(n = 18)	(n = 18)	(n = 18)	(n = 54)

Panel B: Contrast analyses of the differences in evaluation scores[b]

Contrast	DF	Mean square	F-stat.	Sign.
Financial versus non-financial (F/NF)	1	1700.11	7.21	0.0081***
Organization (ORG)	1	264.06	1.12	0.2917
F/NF * ORG	1	697.69	2.96	0.0876*

Panel C: Contrast estimates (effect size) for different subsets[c]



	Effect size	t-stat. (sign.)
Effect of organization		
(1) ORG on financial measures	10.33	2.02 (0.045)**
(2) ORG on non-financial measures	0.17	0.06 (0.955)
Effect financial versus non-financial		
(3) F/NF in unformatted SC	2.85	0.68 (0.496)
(4) F/NF in BSC-format	13.02	3.12 (0.002)***

* Significance levels of 10% (two-tailed).
** Significance levels of 5% (two-tailed).
*** Significance levels of 1% (two-tailed).

[a] Means of differences in evaluation scores (respectively, Std. dev. and number of participants) are shown per cell. We also show the overall means of the non-financial categories, as theory and tests focus on this distinction.

[b] The contrast analyses contrast the performance differences in the financial category against the three other levels with performance differences located in the non-financial categories. The contrast code for F/NF is {3 −1 −1 −1}. The contrast code for organization is {−1 1}. The factor F/NF*ORG, with contrast code {3 −1 −1 −1 −3 1 1 1}, explores whether differences in evaluation scores resulting from performance differences on financial as opposed to non-financial measures depend on how a scorecard is organized of (test of H1).

[c] Panel C explores the effects of organization for performance differences either located on financial measures (1) or non-financial measures (2). We also compare the differences in the weighting for the unformatted scorecard (3) or the BSC-format (4) when performance differences are located on financial or non-financial measures. Effect sizes are derived via contrast estimates (i.e., cells not under consideration are set to 0).

they used in their evaluations. When analyzing these measures, it is important to note that more came from the financial and customer categories (34.7% and 39.7%, respectively) than from the internal business and learning and growth categories (12.7% and 13%, respectively). We also observed that in 81% of the cases, participants listed two or fewer measures from the category that was manipulated (i.e., contained the performance differences).

Given the above, it is important to study where these measures from the manipulated category fall in our participants' list of five measures, in order to explore how participants came to their judgment.[13] If measures of the manipulated category are on the first positions (first two positions) they should get a positive weight. If they are in

[13] Besides those measures from the manipulated category of which, in many cases, only two or fewer are li0sted, participants often supplement their list with customer or financial measures. Therefore, it is important to explore whether measures from the manipulated category appear at the top of the list (in which case, judgment is heavily influenced by measures of the manipulated category) or at the bottom (in which case measures other than those from the manipulated category have a strong influence on judgment).

the back end of the list (last two positions), then we can assume that participants have given more weight to other measures (measures not from the manipulated category) and hence measures of the manipulated category are given a negative weight. [14] We developed a test score which applied this weighting scheme. The more positive the score, the more likely that measures of the manipulated category appear on the first positions of their list. This test score is strongly correlated with the differences in evaluation scores presented in Table 2 (i.e., $r = 0.32$, $p < 0.01$). Analysis of this test score provided further support for our main findings. When performance differences are located in the financial category, participants using the BSC placed the financial measures more upfront than participants using the unformatted scorecard (score of 1.39 versus 0.50, $p < 0.09$). When performance differences are located on the non-financial categories, we found no significant differences in how users of a BSC versus users of the unformatted scorecard listed the relevant measures (-0.11 versus 0.28, $p > 0.20$). In sum, consistent with H1, users of a BSC-format focus more heavily on financial measures than users of an unformatted scorecard. A further test looked at the positions of financial measures, when performance was manipulated on the non-financial categories. It confirmed that even in such cases BSC users put financials more upfront than users of the unformatted scorecard (0.78 versus -0.037; $p < 0.02$) even though the financial category was not manipulated.

Experiment 2

Experiment 2 is similar in design to Experiment 1, except that we added markers to the performance measures in both types of scorecards. We test our second hypothesis, which predicts a main effect of information organization: when scorecards contain markers, users of a BSC, as opposed to those using an unformatted scorecard, will place more weight on measures from both the financial and non-financial categories.

Experimental manipulations, participants, and procedures

In Experiment 2, we again study both the effects of how information is organized (i.e., in an unformatted scorecard versus in a BSC) and the 'type of measure' manipulation (i.e., performance differences between the two businesses units are located in one of the BSC's four different categories). This time we supplement the scorecards with +, −, or = signs (i.e., performance markers) to indicate above-target, below-target, or on-target performance. Fig. 1 presents the screenshots of the marked BSC condition as displayed to our participants. The presentation order of Streetware and Family Fashion was again counterbalanced. A total of 144 students participated in Experiment 2, none of whom had participated in Experiment 1. Nevertheless, because participants in Experiment 2 were recruited from similar courses as those in Experiment 1, no significant differences in the participants' demographics existed across experiments. Participants in Experiment 2 had a mean level of 3.79 years of part-time work experience, and 62.5% were male. Most (97.9%) had visited a clothing store in the past 12 months. As in Experiment 1, participants were randomly assigned to the between-subjects conditions. Sessions lasted about one hour.

Results

Consistent with H2, the means in Table 3, Panel A show that the presence of performance markers causes evaluators to weight both financial and non-financial measures (in particular in the customer and learning and growth perspective) more heavily when presented in a BSC than in an unformatted scorecard. As predicted by H2, Panel B of Table 3 shows a strong main effect of information organization ($F = 8.59$, $p < 0.01$). There is only a weak main effect of difference in weighting of financial versus the non-financial measures ($F = 3.33$, $p = 0.07$). The interaction is not significant ($F = 0.56$, $p = 0.45$).

Panel C of Table 3 also shows that organizing information into a BSC strongly affects evaluations both when performance differences are located in the financial category (+11.06, $p = 0.036$) as well as in the non-financial categories (+6.52, $p = 0.032$).[15] Because both financial and non-financial measures are weighted more heavily, the difference in how financial versus non-financial measures are weighted in a marked BSC is only marginally significant (+7.77, $p = 0.07$). Also, this difference in weighting does not vary between the BSC and the unformatted scorecard because the effect size of this interaction is equal to 4.53 ($= 7.77 - 3.24$) and not significant ($p = 0.45$). Hence, in Experiment 2, we do not find that organizing measures into the BSC intensifies the tendency to weight financial measures more heavily than non-financial measures, as was the case in Experiment 1. In sum, when performance markers are used, organizing measures into the BSC causes evaluators to increase the weight on measures of any category containing a consistent performance difference. As predicted, markers help evaluators to focus on those categories containing consistent performance differences because these markers cause these categories to stand out relative to other categories in the scorecard. [16]

[14] The first two positions receive a weight of 1, the middle position a weight of zero, and the last two positions a weight of −1. Alternative weightings of measures, such as {2, 1, 0, −1, −2}, produced similar results, with the exception that these results are significant on a one-tailed level.

[15] Also here we looked at what happened at the individual business-unit level. The fact that non-financials increase in weight is due to Family Fashion. Here as a result of markers, people weight the negative performance on non-financials more heavily under a BSC than under the unformatted scorecard (51.96 versus 44.46, $p < 0.01$). For financials, people with the BSC gave more weight to the positive performance of Streetware than users of the unformatted scorecard (75.06 versus 65.83, $p < 0.01$). Overall, combining these two effects, people with a marked BSC (relative to users of the marked unformatted scorecard) increase their weighting of performance differences located on financial as well as on non-financial categories.

[16] Again, we find no significant differences in variability between the BSC and the unformatted scorecard. The standard deviation did not vary significantly between these two groups (p-value of the Levene's test = 0.52). Also, the standard deviations of cell means on financials (18.49 versus 11.62, p-value of the Levene's test = 0.19) and on non-financials (17.59 versus 17.08, p-value Levene's test = 0.95) did not differ significantly between the BSC and the unformatted scorecard. Also, across experiments, variability of scores in Experiment 1 and Experiment 2 were not significantly different (p-value of Levene's test = 0.45).

Business unit: Streetware — Year: 20XX

Financial	TARGET	ACTUAL	+/-
Sales margins	60.0%	66.0%	+
Sales growth per store	15.0%	15.8%	+
Inventory turnover	6.0	6.6	+
Percentage of sales from new stores	30.0%	31.5%	+
Customer-related			
Customer satisfaction rating	85.0%	93.5%	+
Sales per square foot of retail space	30000	30000	=
Repeat sales	30.0%	27%	-
# new items in which first to market	70.0	70.0	=
Internal Business Processes			
Returns to suppliers	6.0%	5.4%	+
Average markdowns in percent	15.0%	15.0%	=
Orders filled within one week	3000	2700	-
# stock-outs	2.0	2.0	=
Learning and growth (L&G)			
Hours of sales training per employee	15.0	16.5	+
Employee suggestions per employee	1.0	0.9	-
Retail experience of sales managers	3.0	3.0	=
Employee satisfaction	80.0%	80.0%	=

Business unit: Family Fashion — Year: 20XX

Financial	TARGET	ACTUAL	+/-
Sales margins	62.0%	55.8%	-
Sales growth per store	18.0%	17.1%	-
Inventory turnover	5.0	4.5	-
Percentage of sales from new stores	25.0%	23.8%	-
Customer-related			
Customer satisfaction rating	90.0%	99%	+
Sales per square foot of retail space	25000	25000	=
Repeat sales	40.0%	36%	-
# new items in which first to market	60.0	60.0	=
Internal Business Processes			
Returns to suppliers	4.0%	3.6%	+
Average markdowns in percent	12.0%	12.0%	=
Orders filled within one week	2500	2250	-
# stock-outs	3.0	3.0	=
Learning and growth (L&G)			
Hours of sales training per employee	13.0	14.3	+
Employee suggestions per employee	2.0	1.8	-
Retail experience of sales managers	4.0	4.0	=
Employee satisfaction	82.0%	82.0%	=

Fig. 1. The marked BSC screenshots used in Experiment 2 for the manipulation in which performance differences are located on the financial measures. The scorecard used in Experiment 1 is identical except that the performance markers (i.e., the +, −, and = signs) are not displayed. When administering the cases, we randomized which measures were on-target, below-target, or above-target such that two of the four measures in the category containing the performance differences were 5% above target for Streetware (below target for Family Fashion) while the other two measures were 10% above target for Streetware (below target for Family Fashion). In the category with no performance differences between business units, one measure was 10% above target, one 10% below target, and two on target (see Footnote 6). (1) Screenshots for the marked BSC (performance difference on financial measures).

Supplementary analyses

Also in Experiment 2, participants listed more measures from the financial and customer category (37.6% respectively 38.1%) than from the internal business and learning and growth category (11.5% versus 12.8%) after their evaluation judgment. Again, in more than 77% of the cases only two or fewer measures from the manipulated category were listed. When applying the same test score as used in Experiment 1, we again find support for our arguments. Again, the score strongly correlates with the differences in evaluation scores as analyzed in Table 3 ($r = 0.25$, $p < 0.01$). Markers cause participants using a BSC to give more weight to both financial and non-financial measures. When performance differences were located in the financial category, participants using the BSC placed the financial measures more upfront than participants using the unformatted scorecard (score of 1.67 versus 0.39, $p < 0.01$). When performance differences were located in the non-financial categories, BSC users placed the relevant measures (i.e., those from the manipulated non-financial category) higher on their lists than users of the unformatted scorecard (0.20 versus −0.30, $p = 0.095$). Hence, consistent with H2, supplementing the BSC with markers ensures that evaluators place more weight on the measures of the category containing the performance differences.

Effects of adding markers

For explorative reasons, we make a comparison between Experiments 1 and 2, to examine the effect of adding markers to the scorecards. Given that we are comparing across experiments, the statistical results should be treated with caution. As Table 4 shows, adding markers to a BSC increases the weight given to non-financials more strongly than when markers are added to an unformatted scorecard (12.47 versus 6.11 = +6.36, $p = 0.07$). We, however, observe no difference in how the addition of markers to a BSC versus to an unformatted scorecard affects the weighting of measures in the financial category (+7.22 − 6.50 = 0.72, p is ns). This can explain the main effect in Experiment 2 of information organization. With markers, financials receive a similar increase in weights in a BSC as they do in an unformatted scorecard. As a result, financials (weighted more heavily in Experiment 1), continue to be more heavily weighted in a BSC in Experiment 2. Because markers have a greater impact on how the non-financial measures are weighted in a BSC than in an unformatted scorecard, our non-financials (weighted the same in either type of scorecard in Experiment 1), are weighted in Experiment 2 more heavily in a BSC. Moreover, Table 4 further shows that non-financials get the largest weight in a marked BSC (24.12) than in the other three scorecards (17.59, 11.65, and 11.48). This disordinal contrast is significant (10.54, $p < 0.01$, results not tabulated). These results likely have practical implications. If firms want evaluators to pay more attention to their non-financial measures, one approach is organizing the measures into a marked BSC.

Discussion

Our paper studies how variations in the format of scorecards and the presentation of measures therein affect how evaluators weight financial versus non-financial information in performance evaluations. Experiment 1 shows that when performance differences are located in the financial category, BSC users place more weight on financial

Table 3
Results of Experiment 2 (test of H2).

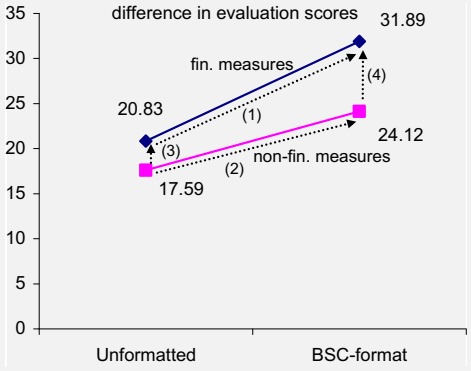
Panel A: Summary statistics per experimental cell[a]

Scorecard organization	Type of measure				
	Financial measures	Customer	Internal	L&G	Non-financial measures
Unformatted scorecard with markers					
Eval. Streetware	65.83	73.28	71.28	64.11	69.56
Eval. Family Fashion	45.00	45.61	53.89	56.39	51.96
Difference in eval.	20.83	27.67	17.39	7.72	17.59
	[18.49]	[15.46]	[15.95]	[16.13]	[17.59]
	(n = 18)	(n = 18)	(n = 18)	(n = 18)	(n = 54)
BSC-format with markers					
Eval. Streetware	75.06	73.33	62.45	69.94	68.57
Eval. Family Fashion	43.17	39.44	46.43	47.50	44.46
Difference in eval.	31.89	33.89	16.02	22.44	24.12
	[11.62]	[12.76]	[18.75]	[14.89]	[17.08]
	(n = 18)	(n = 18)	(n = 18)	(n = 18)	(n = 54)

Panel B: Contrast analyses for differences in evaluation scores[b]

Contrast	DF	Mean square	*F*-stat.	Sign.
Financial versus non-financial (F/NF)	1	818.68	3.33	0.0702*
Organization (ORG)	1	2110.64	8.59	0.0040***
F/NF * ORG	1	138.61	0.56	0.4540

Panel C: Contrast estimates (effect size) for different subsets[c]



	Effect size	*t*-stat. (sign.)
Effect of organization		
(1) ORG on financial measures	11.06	2.12 (0.036)**
(2) ORG on non-financial measures	6.52	2.16 (0.032)**
Effect financial versus non-financial		
(3) F/NF in unformatted SC	3.24	0.76 (0.449)
(4) F/NF in BSC	7.77	1.82 (0.071)*

* Significance levels of 10% (two-tailed).
** Significance levels of 5% (two-tailed).
*** Significance levels of 1% (two-tailed).

[a] Means of differences in evaluation scores (respectively, Std. dev. and number of participants) are shown per cell. We also show the overall means of the non-financial categories, as theory and tests focus on this distinction.

[b] Contrast analyses and estimates in Experiment 2 are equivalent to the contrast analysis and estimates performed in Table 2 for Experiment 1 (refer to Table 2 for more detail on the contrast codes). Given our prediction in H2, we predict that organization has a strong effect regardless of the type of measures containing the performance differences; hence, we do not presume a significant interaction of F/NF * ORG.

[c] Similar subset analyses as used in Table 2, Panel C. The number attached to each arrow in the figure refers to the number of subset analyses displayed on the figure's right-hand side.

measures than do users of an unformatted scorecard. In contrast, when performance differences are located in one of the non-financial categories, the type of scorecard used (i.e., a BSC versus an unformatted scorecard) does not affect performance evaluations. Experiment 2, however, demonstrates that with the addition of performance markers, organizing measures into a BSC increases the weight evaluators attach to performance differences located on both financial and non-financial measures. Ultimately, performance differences on non-financial measures, receive the greatest weight in evaluations when presented in a marked BSC.

We extend the results of Lipe and Salterio (2002) in two important ways. First, we show that organizing information in a BSC compared to in an unformatted scorecard can increase (rather than decrease) the weight evaluators attach to a particular category of performance measures, especially when performance differences are located in the financial category. A BSC simplifies the task of identifying the financial measures and assessing them in combination, can reinforce the evaluator's tendency to rely more on the financial measures. Second, in Experiment 2, we show that, when we add performance markers to the scorecards, a BSC can increase an evaluator's attention toward any

Table 4
Comparison between Experiment 1 and Experiment 2.

	FIN	NFIN
Unformatted scorecard (Experiment 1)	14.33	11.48
Marked unformatted SC (Experiment 2)	20.83	17.59
Effect marker	+6.50	+6.11
	FIN	NFIN
BSC-format (Experiment 1)	24.67	11.65
Marked BSC-format (Experiment 2)	31.89	24.12
Effect marker	+7.22	+12.47
	0.72 ($p = 0.46$)	.36 ($p = 0.07$)*

The table compares Experiment 1 and Experiment 2 to investigate the effect of adding markers to the scorecard. We program the effects of the comparison using contrast estimates. Results are based on a one-sided *t*-test, given that markers will increase the weighting of performance differences.

* Significance levels of 10% (one-tailed).

type of category therein that contains a performance difference, be it financial or non-financial.

Our findings have important practical implications. Some firms use a BSC to emphasize the leading non-financial indicators of firm value. Subtle changes in the presentation of information in a BSC (such as adding performance markers) can offer a solution to firms who want to use a BSC to increase the weight evaluators assign to such indicators of firm value. Without performance markers, business-unit managers may react negatively to the use of a BSC for fear that evaluators will not fully incorporate these non-financials into their evaluations (see Ittner et al., 2003; Malina & Selto, 2001).

Our study also offers some opportunities for further research. First, prior studies (e.g., Banker et al., 2004; Lipe & Salterio, 2000) have shown that evaluators favor common and general measures over unique and strategy-linked measures. One important suggestion for studies that focus on this problem of common–unique measures is to explore whether unique non-financial measures are more easily ignored than unique financial measures in a BSC-format, because evaluators tend to focus more strongly on financial measures when measures are organized in a BSC-format.

Second, while our experiment employed students who had received instruction in the BSC, it would be interesting to explore how certain presentation features in a BSC affect more experienced managers, whose knowledge of, for example, measurement properties and causal relationships across measures is more developed (Bonner & Lewis, 1990). This might cause them to focus less intensely on financials. Prior work has, however, shown that experienced managers also face cognitive processing limitations (Shanteau, 1988, 1992) similar to less knowledgeable evaluators (Dilla & Steinbart, 2005a). Simple changes to the presentation of information, like performance markers, might therefore also help them to better deal with a large set of measures.

Third, we located similar performance differences between two business units in each of the four BSC perspectives. Future work, however, can study how participants weight performance information when the business units themselves are less distinguishable on a specific BSC category. For example, one business unit might score well in the financial category, whereas the other might score well on a non-financial category. In addition, one might spread excellent performance across multiple categories. It is interesting to then study how different presentation formats facilitate the processing of performance information.

Fourth, the weights evaluators attach to different types of performance measures may well depend on strategy (as well as the information provided about that strategy) and other factors in the operating environment (see e.g., Banker et al., 2004; Lillis & van Veen-Dirks, 2008; van Veen-Dirks, 2006, 2010). Future research can disentangle how information about such factors interacts with the organization and presentation of performance measures.

Finally, researchers can explore the use of other presentation features, such as graphs, traffic lights, or aggregations of measures in formulas (Cardinaels, 2008; Dilla & Steinbart, 2005b; Roberts et al., 2004). Certainly, if a particular firm has derived a set of measures that are known to drive firm value, it is important that evaluators use these measures in their evaluations and, consequently, that business-unit managers use these measures in their daily decisions (Feltham & Xie, 1994; Holmstrom & Milgrom, 1991). We therefore support continued research into how different types of scorecards, as well as other factors in the evaluation process, inhibit or stimulate such use.

Acknowledgements

We want to thank Mike Shields (editor) and the two anonymous referees for their helpful suggestions. We further want to thank Maggie Abernethy, Jan Bouwens, Penelope Cray, Chris Ittner, Ken Merchant, Mina Pizzini, Steve Salterio, Ed Vosselman, William Waller, and seminar participants at Tilburg University, the University of Leuven, the ARN and ERIM seminars in Rotterdam, the MAS mid-year Conference in Tampa, the EIASM conference for new directions in management accounting in Brussels, and the GMARS conference in Sydney for their helpful comments.

References

Almer, E. D., Hopper, J. R., & Kaplan, S. E. (2003). A research tool to increase attention to experimental materials: Manipulating presentation format. *Journal of Business and Psychology, 17*(3), 405–418.

Anthony, R. N., & Govindarajan, V. (2001). *Management control systems* (10th ed.). New York: McGraw-Hill.

Azofra, V., Prietro, B., & Santidrian, A. (2003). The usefulness of a performance measurement system in the daily life of an organisation: A note on a case study. *British Accounting Review, 35*, 367–384.

Baddeley, A. (1994). The magical number seven: Still magic after all these years. *Psychological Review, 101*(2), 353–356.

Banker, R. D., Chang, H., & Pizzini, M. J. (2004). The Balanced Scorecard: Judgmental effects of performance measures linked to strategy. *The Accounting Review, 79*(1), 1–23.

Bauer, J., Tanner, S. J., & Neely, A. (2004). Developing a performance measurement audit template—A benchmarking study. *Measuring Business Excellence, 8*(4), 17–25.

Bonner, S. E., & Lewis, B. L. (1990). Determinants of auditor expertise. *Journal of Accounting Research, 28*, 1–19.

Buckless, F. A., & Ravenscroft, S. P. (1990). Contrast Coding: A refinement of ANOVA in behavioral analysis. *The Accounting Review, 65*(4), 933–945.

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in information visualization: Using vision to think*. San Diego: Academic Press.

Cardinaels, E. (2008). The interplay between cost accounting knowledge and presentation formats in cost-based decision making. *Accounting, Organizations and Society, 33*(6), 582–602.

DeBusk, G. K., Brown, R. M., & Killough, L. N. (2003). Components and relative weights in utilization of performance measurement systems like the Balanced Scorecard. *British Accounting Review, 35*(3), 215–231.

Dilla, W. N., & Steinbart, P. J. (2005a). Relative weighting of common and unique Balanced Scorecard measures by knowledgeable decision makers. *Behavioral Research in Accounting, 17*, 43–53.

Dilla, W. N., & Steinbart, P. J. (2005b). The effects of alternative supplementary display formats on Balanced Scorecard judgments. *International Journal of Accounting Information Systems, 6*, 159–176.

Feltham, G. A., & Xie, J. (1994). Performance measure congruity and diversity in multi-task principal/agent relations. *The Accounting Review, 69*(3), 429–453.

Few, S. (2006). *Information dashboard design: The effective visual communication of data*. Schastopol: O'Reilly Media, Inc..

Fiske, S. T., & Taylor, S. E. (2008). *Social cognition: From brains to culture*. Boston: McGraw-Hill Higher Education.

Ganzach, Y. (1994). Feedback representation and prediction strategies. *Organizational Behavior and Human Decision Processes, 59*, 391–409.

Ghosh, D., & Lusch, R. F. (2000). Outcome effect, controllability, and performance evaluation of managers: Some field evidence from multi-outlet business. *Accounting, Organization and Society, 25*(4–5), 411–425.

Hastie, R. (1991). A review from a high place: The field of judgment and decision making as revealed in current textbooks. *Psychological Science, 2*, 135–138.

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin, 17*(3), 311–327.

Haynes, C. M., & Kachelmeier, S. J. (1998). The effects of accounting contexts on accounting decisions: A synthesis of cognitive and economic perspectives in accounting experimentation. *Journal of Accounting Literature, 17*, 97–136.

Holmstrom, B., & Milgrom, P. (1991). Multitask principle agent analyses: Incentive contracts, asset ownership and job design. *Journal of Law, Economics, and Organization, 7*, 24–52.

Hoque, Z., & James, W. (2000). Linking Balanced Scorecard measures to size and market factors: Impact on organizational performance. *Journal of Management Accounting Research, 12*, 1–17.

Ittner, C. D., & Larcker, D. F. (1998). Innovations in performance measurement: Trends and research implications. *Journal of Management Accounting Research, 10*, 205–238.

Ittner, C., Larcker, D. F., & Meyer, M. W. (2003). Subjectivity and the weighting of performance measures: Evidence from a Balanced Scorecard. *The Accounting Review, 78*(3), 725–758.

Kaplan, R. S., & Norton, D. P. (1992). The Balanced Scorecard: Measures that drive performance. *Harvard Business Review, 70*(1), 71–79.

Kaplan, R. S., & Norton, D. P. (1996). *The balanced scorecard: Translating strategy into action*. Boston: Harvard Business School Press.

Kaplan, S. E., & Wisner, P. S. (2009). The judgmental effects of management communications and a fifth Balanced Scorecard category on performance evaluation. *Behavioral Research in Accounting, 21*(2), 37–56.

Kleinmuntz, D. N., & Schkade, D. A. (1993). Information displays and decision processes. *Psychological Science, 4*(4), 221–227.

Koonce, L., Lipe, M. G., & McAnnally, M. L. (2005). Judging the risk of financial instruments: Problems and potential remedies. *The Accounting Review, 80*(3), 871–895.

Kulik, C. T. (1989). The effects of job categorization on judgments of the motivating potential of jobs. *Administration Science Quarterly, 34*(1), 68–90.

Libby, R. (1981). *Accounting and human information processing*. Englewood Cliffs, NJ: Prentice Hall.

Libby, T., Salterio, S. E., & Webb, A. (2004). The Balanced Scorecard: The effects of assurance and process accountability on managerial judgment. *The Accounting Review, 79*(4), 1075–1094.

Liedka, S. L., Church, B. K., & Ray, M. R. (2008). Performance variability, ambiguity intolerance, and Balanced Scorecard-based performance assessments. *Behavioral Research in Accounting, 20*(2), 73–88.

Lillis, A. M., & van Veen-Dirks, P. M. G. (2008). Performance measurement system design. *Journal of Management Accounting Research, 20*, 25–57.

Lipe, M. G., & Salterio, S. (2000). The Balanced Scorecard: Judgmental effects of common and unique performance measures. *The Accounting Review, 75*(3), 283–298.

Lipe, M. G., & Salterio, S. (2002). A note on the judgmental effects of the Balanced Scorecard's information organization. *Accounting, Organizations and Society, 27*(6), 531–540.

Lohman, C., Fortuin, L., & Wouters, M. (2004). Designing a performance measurement system: A case study. *European Journal of Operational Research, 156*, 267–286.

Luft, J., & Shields, M. (2001). The effects of financial and nonfinancial performance measures on judgment and decision performance. Working paper, Michigan State University.

Maines, L. A., & McDaniel, L. S. (2000). Effects of comprehensive-income characteristics on non-professional investors' judgments: The role of financial-statement presentation format. *The Accounting Review, 75*(2), 179–207.

Malina, M. A., Norreklit, H. O., & Selto, F. H. (2007). Relations among measures, climate of control, and performance measurement models. *Contemporary Accounting Research, 24*(3), 935–982.

Malina, M. A., & Selto, F. H. (2001). Communicating and controlling strategy: An empirical study of the effectiveness of the Balanced Scorecard. *Journal of Management Accounting Research, 13*, 47–90.

Merchant, K. A., & Van der Stede, W. A. (2007). *Management control systems: Performance measurement, evaluation and incentives* (2nd ed.). Prentice Hall.

Mitchell, T., & Kalb, L. (1981). Effect of outcome knowledge and outcome valence on supervisors' evaluations. *Journal of Applied Psychology, 6*(1981), 604–612.

Payne, J. W. (1982). Contingent decision behavior. *Psychological Bulletin, 92*, 382–402.

Payne, J. W., Bettman, J. R., & Luce, M. F. (1998). Behavioral decision research: An overview. In M. H. Birnbaum (Ed.), *Measurement, judgment, and decision making*. San Diego: Academic Press.

Roberts, M. L., Albright, T. L., & Hibbets, A. R. (2004). Debiasing Balanced Scorecard evaluations. *Behavioral Research in Accounting, 16*, 75–88.

Schiff, A. D., & Hoffman, L. R. (1996). An exploration of the use of financial and nonfinancial measures of performance by executives in a service organization. *Behavioral Research in Accounting, 8*, 134–153.

Schkade, D. A., & Kleinmuntz, D. N. (1994). Information displays and choice processes: Differential effects of organization, form, and sequence. *Organizational Behavior and Human Decision Processes, 57*(3), 319–337.

Shanteau, J. (1988). Psychological characteristics and strategies of expert decision makers. *Acta Psychologica, 68*, 203–215.

Shanteau, J. (1992). How much information does an expert use? Is it relevant. *Acta Psychologica, 81*, 75–86.

Silver, M. S. (1991). Decisional guidance for computer-based decision support. *MIS Quarterly, 15*(1), 105–122.

Smith, M. J. (2002). Gaming nonfinancial performance measures. *Journal of Management Accounting Research, 16*, 183–205.

Sundstrom, G. A. (1987). Information search and decision making: The effects of information displays. *Acta Psychologica, 65*, 165–179.

Van Veen-Dirks, P. M. G. (2006). Complementary choices and management control: Field research in a flexible production environment. *Management Accounting Research, 17*, 72–105.

Van Veen-Dirks, P. M. G. (2010). Different uses of performance measures: The evaluation versus reward of production managers. *Accounting, Organizations and Society, 23*(2), 141–164.

Vera-Muñoz, S. C., Kinney, W. R., & Bonner, S. E. (2001). The effects of domain experience and task presentation format on accountants' information relevance assurance. *The Accounting Review, 76*(3), 405–429.